Migration Letters

Volume: 20, No: S13(2023), pp. 435-448

ISSN: 1741-8984 (Print) ISSN: 1741-8992 (Online)
www.migrationletters.com

Deep Learning Based Human Emotion Exposure Detection Using Vocal and Countenance

Dr.C. Pabitha[1] and Dr.B. Vanathi[2]

Abstract

*The subtle expression of emotions on a person's face might give insight into the ideas running through that person's head. Today's society depends on being able to read facial emotions. Faces are a universal language that humans use to communicate a common and basic set of emotions. In a variety of settings and spheres of life, emotion recognition has benefits. For the purposes of safety and health, it is beneficial and essential. Additionally, it is essential for quickly and simply determining human emotions at a particular time without actually asking. Facial expression with non verbal cues are important for interpersonal interaction. The method of determining a person's emotional state is done using a facial expression recognition system. The technique of identifying different facets of a person's facial expressions is called emotion recognition. One of the most potent and challenging jobs in social communication is identifying human emotion from video frames, or facial expressions. This system compares the captured image to the training dataset that is stored in the database, and then displays the emotional state of the image. Recognizing automatically, the human emotions in pictures and videos will be possible with the help of an algorithm that carries out the identification, extracting and evaluating these facial expressions. These systems allow detection for emotions such as joy, rage, sadness, disgust, surprise, fear, neutral, and more. The applications for DL-based human emotion detection attempt to comprehend the significance of these features of body language and apply this understanding to various datasets and sources of information for the face emotion CNN pretrained module mobilenetv2 employed. This can assist in identifying and extracting facial expressions. These algorithms outperform numerous datasets of photos and videos in terms of accuracy***.***

Keywords: *Emotion Recognition, CNN, Deep Learning Vocal Detection.*

Introduction

Our daily decisions are influenced by emotions because they play a significant role in the human experience and rule our daily lives. When something makes us happy, we often repeat it, but when something makes us angry or depressed, we avoid it. While analyzing psychological theory, we can define emotion as "A severe state of psychological phase including three different components: a subjective experience, a physiological response, and a behavioral or expressive response." This explanation conveys two concepts to us:

a. Subjectivity and complexity: Different people respond differently to the same circumstances. What could be humorous to a person will be offensive to another. In addition, we offer a variety of emotions at once, making it challenging to comprehend our own feelings, let alone those of another.

[1] Assistant Professor, Department of Computer Science and Engineering, SRM Valliammai Engineering College. pabithac.cse@srmvalliammai.ac.in

[2] Professor and Head, Department of Computer Science and Engineering, SRM Valliammai Engineering College. vanathib.cse@srmvalliammai.ac.in

b. Response: Emotions are accompanied by a physical reaction, such as perspiration, an accelerated heartbeat, or increased respiration, similarly gestures and expressions displayed on face like crossing arms, frowning etc..

Emotion Recognition

When interacting and socializing with other individuals, emotions are exhibited. It can be difficult to study how to understand them, thus technology is utilized to accomplish that. Today, a lot of institutions use emotion recognizing technology. The technique of finding human emotion displayed by both verbal and facial responses is known as emotion recognition. As we've seen, NLP approaches, deep learning and linguistics are utilized to identify emotions in text. Given that subjectivity in language and phenomena like irony and sarcasm make sentiment analysis difficult, emotion recognition attempts to provide a more thorough picture of a person's reactions. But frequently, sentiment analysis and emotion detection are misunderstood, so let's delve a little deeper. While sentiments expressed in the text is neutral, negative, or positive, sentiment analysis typically derives a polarity from the text. Although very helpful, this does not explore the fundamental causes of the sentiment output. With emotion recognition, we want to identify this additional analysis. Emotion recognition is considered to be the mail facial recognition systems that have evolved and expanded throughout time. Because of the use of facial emotion recognition software, a specific computer can now analyze and process a person's facial emotions. This application uses sophisticated picture dispensation to imitate the way a brain of human works, permitting it to discern emotions also. Artificial intelligence (AI) or "artificial intelligence" identifies and analyses diverse facial emotions in order to utilise them in conjunction with other information provided to it. This enables authorities to understand a person's emotions using only technology, which is necessary for a variety of duties such as investigations and interviews. Different facial expressions can be identified and recognized using emotion recognition software, which uses facial expression analysis to do so. Deep learning is mainly considered as a subset of the broad field termed as Artificial Intelligence for facial recognition that mimics how the brain of a human develops patterns for item detection and even decision-making by processing data. It is very prominent in research domain and considered to be the subset of the field AI and machine learning technology. Neural net is the foundation of deep learning. An algorithm called a neural net is modelled after the cerebral cortex and works similarly to the brain. Because of the advantages it possesses for emotion recognition, it has grown in popularity over time and especially recently. The first layer is input, secondly hidden layer, and lastly the output layer of the neural network are three layers, precisely like the cerebral cortex. In the neural network, preferred data can be added, and it all passes through these layers. Each layer alters each input value in an effort to produce the desired and target output. For the purpose of ensuring accurate and acceptable outputs, the emotion detection software goes through training. For algorithms to work, inputs and outputs must be understood. The algorithms must therefore be able to detect human expressions. There are two methods to accomplish this possibility.

a. Categorical: According to this method, there is a finite set of emotions that fall into different set classes. Contempt, grief, shock, disgust, wrath, happiness, and fear are among the emotions that are mentioned.
b. Dimensional: This method assumes that emotions exist on a spectrum and cannot be concretely described. The Circumflex model of affect has two dimensions, but the PAD emotional state includes three.

When modelling them using machine learning, the approach choice has some significant ramifications. In case, a categorical model is chosen for a human emotion, classifier is built, and labels and pictures are titled with emotions of human such as "happy," "sad," and so on. But if we use a model of different dimensions then the outputs should be on a sliding scale. There are other inputs that can be employed or provided to the system for analysis, therefore selecting a model or set of training data is never considered to be the end of the process. The primary areas for emotion recognition are listed below:

a. Video: This is the major datasets used in emotion-based computing, and much work is still being done to perceive how effectively to use them for emotion recognition. The cameras on mobile phones are an example of contemporary software for this industry.
b. Image: It is one of the major sets that are available for emotion recognition, much like video. According to certain studies, video sequences may be categorized into different genres and visuals might be automatically tagged with various emotions.
c. Speech: Normally, talks are transcribed into texts so that they may be analyzed, however this approach would not work for recognizing emotions. Research is still being conducted to examine whether speech can be used to recognize emotions instead of transcribed texts.
d. Text: DTM also termed as Document Team Matrix is the most acknowledged data structure in Text used widely. It is a matrix in structure which includes different records indicating the frequency of Text/ words in a document. However, because it uses individual words, it is ineffective for determining emotions. A text can be perceived depending on its nature, punctuation etc, and these elements are all taken into account by analyst in their continual development of novel text data structures.

Emotion recognition mainly focuses on gathering emotions from conversations between two or more people. Emotion based Datasets in this category are often free samples taken from many of the social media platforms. Still, many obstacles remain in this subject, such as the availability of sarcasm in a discourse, and the way the person conversing shifts in mood and the abrupt change in context keep on changing. Nowadays, emotion recognition is employed for a variety of applications that many people are unaware of on a daily basis. Here are some examples of how emotion awareness can be beneficial:

a. Security Measures: Recognizing Expressions or emotions is already used in educational sectors and other institutions to avoid any illegal activities and violence in any form.
b. HR Assistance: Some businesses employ Artificial Intelligence aided with emotion recognition as an API capability for assisting HR departments. The algorithm aids in identifying the genuineness of the worker in doing the job by analyzing his facial expressions and intentions.
c. Customer Service: AI induced system that automatically recognize customer satisfaction of customer are placed in many systems service centre. The expression of customer is tracked before entering the Service centre and after getting serviced. If the customer expression is not happy, the rating of service is sent to the employer for taking further actions.
d. Differently Abled youngsters: A project is underway that will use a system in Google Glass smart glasses to help autism child and also in understanding the emotions of all person in the environment. Whenever a person interacts with another, the images convey the feelings of other person’s emotions..
e. Audience Engagement: Organizations are also employing emotion recognition to predict commercial results based on the emotional responses of their audiences. Apple also introduced Animoji, a new feature in iPhones that allows an emoji to replicate a person's facial emotions.
f. Video Game Testing: Video games are evaluated in order to get user feedback and determine whether or not corporations achieve their aims. During these testing phases, emotion recognition can be used to interpret a user's real-time feelings, and their feedback can be incorporated into the final product.
g. Healthcare: Industry like healthcare is definitely making use of face emotion recognition these days. This aids in patient to make decision as to whether medication itself cures the ailment or physician diagnosis is required for further analysis.
h. Image-Based Model: To classify various emotions displayed in facial expressions based on image sequence, two network of deep learning origin is implemented By

combining CNN and LSTM models, the initial network is responsible for capturing temporal reflections while appearing.

i. Speech-Based Model: Given that vocal components constitute for one-third of all human communication, it is obvious that combining voice signals and a facial image can aid in accurate and natural recognition. As a result, by combining an RNN with the FER, we suggest a plausible feature combination that can increase emotion-recognition performance.

There has been substantial research in the topic of facial emotion recognition (FER) during the last three decades, as reported in the literature. Despite a considerable body of work in this field, no comprehensive comparison of classical machine learning (ML) and deep learning (DL) techniques has been conducted. Some previous research, such as Koakowska's review, focused mostly on classical ML methods, whereas Ghayoumi offered a brief overview of DL in FER and Ko et al. did an in-depth evaluation of facial emotion identification systems utilizing visual data. These papers, however, mostly examined the differences between classic ML and DL approaches.

This research work plans to fill this gap by investigating a thorough analysis and comparison of traditional ML and DL techniques in the context of FER. Our contributions are as follows:

a. Our primary objective is to offer a broad understanding of recent research developments and to assist newcomers in grasping the most important components and raising trends within the FER field.
b. We employ multiple common datasets implementing several video sequences and pictures with varying characteristics and purposes to demonstrate the practical application of both ML and DL techniques.
c. We evaluate the resource utilization and accuracy of DL and traditional ML methods in FER. DL-based approaches tend to achieve excellent accuracy but require longer training times and substantial computational resources, including both CPU and GPU power. This has led to the implementation of many FER techniques in embedded systems like Raspberry Pi, Jetson Nano, and mobile devices.
d. Ultimately, our study provides a comprehensive assessment of facial emotion recognition using classic ML and DL algorithms, shedding light on potential research gaps that may be of interest to new researchers entering the field.

Objective

Emotions are a vital component of human communication. Emotions, conduct, and thoughts are all intertwined in such a way that the mix of these characteristics influences how we act and make judgments. As a result, there has been an increase in interest in this area of scientific research in recent years. Automatic emotion recognition can be used to improve performance in a variety of domains. Human-computer interaction, for example, because identifying the emotional state of a computer system's user allows for a more natural, productive, and intelligent relationship. The ultimate goal of this research paper is to recognize human emotions through facial expressions and voice.

Literature Review

Emotions are detected through speech and facial expressions. Krishna Mohan Kudiri, Abas Md Said, and M Yunus Nayan presented "Emotion Detection through Speech and Facial Expressions" as a study. A person can experience an infinite amount of emotions in this study. Based on the facts supplied above, emotions are divided into two categories: joyful feelings and bad emotions. Positive emotions benefit society, whilst negative emotions do not. Identifying unpleasant sensations amid an infinite variety of emotions is incredibly difficult in nature. To solve this issue, psychologists have identified a small number of fundamental emotions that influence the human mind while making decisions: anger, sadness, happiness, fear, disgust, and surprise. With the exception of the basic emotions,

all extra emotions are a synthesis of the basic emotions. In a related paper, Rao et al. developed a bimodal emotion recognition system integrating prosodic and facial data. As visual data qualities, the areas around the eyes and lips were utilised. Likewise, audio acoustic aspects were used. For classification, neural networks were used. Fernando A et al. conducted research on emotion recognition through speech and facial expression. For speech and facial emotions, they used gender and emotion voice analysis (GEVA) with a Bayesian network classifier and gender and emotion face analysis (GEFA) with an SVM network classifier. The final classification-based fusion was carried out using the SVM classifier. GEVA demonstrates prosodic behavior in this context. The following steps were taken to produce these models:

1) Facial expression model: In this scenario, emotion models are built using facial data sourced from the DaFEx data repository. Each category is defined by a set of 29,880 features (calculated as 166 files multiplied by 3 minutes multiplied by 60 seconds). These feature vectors are then inputted into the classifier to generate emotional models.
2) Speech model: In this instance, 29,880 features are utilized for each category. In a manner analogous to the previous case, the feature vectors are input into the classifier.

While implementing our research paper, we implemented a novel technique for identifying human emotions through image sequences and audio features, employing a weighted integration approach. Our method involves the synchronization of speech signals and visual sequences to achieve emotion recognition. To achieve this synchronization, we employ three deep networks, with one of them dedicated to training on image sequences and specializing in detecting variations in facial expressions. Furthermore “facial landmarks are input into an alternate network to capture facial movement. Meanwhile, the speech signals are initially converted into acoustic attributes, which are subsequently employed to align with the visual sequence in the other network." Furthermore, we provide a novel strategy for integrating the models that outperforms prior integrated methods. To validate the proposed procedure, an accuracy test is performed. The performance results indicates that the proposed method shows better prediction than already analyzed research.

"A review and insight into deep learning for facial emotion recognition", Wafa Mellouka and Wahida Handouzia suggested an essay titled "Examining Deep Learning for Facial Emotion Recognition: A Comprehensive Overview and Key Findings”. To achieve a more refined categorization in our study, the process of feature extraction from one face to another presents a challenging and intricate task. In 1978, Ekman and Freisen pioneered the development of the Facial Action Coding System (FACS), a system that characterizes facial movements using Action Units (AUs). The human face is segmented into 46 AUs, with each AU being associated with one or more facial muscles. While automatic Facial Emotion Recognition (FER) has been extensively explored by researchers, it remains a complex procedure due to the unique way each individual expresses their emotions.

Numerous obstacles and challenges in this domain must not be overlooked, including variations in head pose, lighting conditions, age, gender, and backgrounds, as well as occlusions caused by items like sunglasses, scarves, skin conditions, and more. In the paper titled "Utilizing Deep Learning for Facial Emotion Recognition," Mollahosseini and colleagues present a deep Convolutional Neural Network (CNN) method for recognizing emotions in facial expressions across multiple contemporary datasets. Initially, they extract facial landmarks from the dataset and then resize the images to dimensions of 48x48 pixels. To enhance the dataset, they employ data augmentation techniques. The network architecture comprises two convolution-pooling layers followed by two inception-type modules, incorporating convolutional layers of sizes 1x1, 3x3, and 5x5. They also utilize the network-in-network technique to enhance local performance and reduce the risk of overfitting.

In their study, Kim and colleagues explore the time-related changes in facial expressions when transitioning between emotional states. They put forth a spatiotemporal framework that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM). Initially, the CNN is responsible for understanding the spatial characteristics of facial expressions within frames corresponding to emotional states. Subsequently, an LSTM is employed to capture the sequential patterns of these spatial features.

In their work, Yu and colleagues introduce the Spatio-Temporal Convolutional with Nested LSTM (STC-NLSTM) structure, which encompasses three distinct deep learning subsystems: a 3D CNN for extracting spatiotemporal features, a temporal T-LSTM for preserving temporal dynamics, and a convolutional C-LSTM for modeling multi-level features.

In the study titled "Detection of Gender, Age, and Emotion of a Human Image Using Facial Features," Wei-Long Zheng and Bao_x0002_Liang Lu propose a project that employs the Rough Contour Estimation Routine (RCER) neural network technique. They achieve a recognition accuracy of 92.1% by describing both a radial basis function network (RBFN) and a multilayer perceptron (MLP) network. While MLP offers the highest accuracy, it comes with the drawback of a high parameter count due to full connectivity.

The paper "Emotion Recognition Using Component Analysis" by Zixing Zhang, Fabien Ringeval, Eduardo Coutinho, Erik Marchi, and Björn Schüller utilizes Principal Component Analysis (PCA) on the FACE94 dataset. This approach results in a 35% reduction in computing time and a 60% improvement in recognition accuracy. The advantage lies in improved accuracy and reduced processing time, but the authors express the desire to replicate their experiments on larger and more diverse databases.

Lastly, the use of two-dimensional visual data for emotion recognition is explored in A. Yao, D. Cai, P. Hu, S. Wang, L. Shan, and Y. Chen suggested "Emotion recognition using 2d images" as a study. In this study, a method involving 2D Gabor filters applied to random images was employed. Precision in utilizing the 12 Gabor Filters to detect edges was a focal point. The researchers developed a Multichannel Gabor filtering approach for the purpose of identifying salient points and extracting texture information to facilitate image retrieval. Notably, they placed emphasis on incorporating both global and local color histograms, as well as factors related to object shapes within the images. However, it's worth noting that this image processing technique entails extended training times and represents a substantial computational method.

In the paper titled "Emotion recognition using PCA and LDA features," authored by Yelin Kim and Emily Mower Provos within the context of "Emotion recognition using PCA and LDA features," the study employed Local Gabor Filter PCA LDA (JAFFE). The utilization of PCA+LDA features led to an impressive identification rate of 97.33%. The researchers propose and conclude that PCA+LDA features offer a partial solution to the issue of sensitivity to illumination. "The benefit lies in their detailed explanation of the five modules utilized to attain favorable outcomes, which include Palmprint Acquisition, Pre-processing, Textured Feature Extraction, Matching, and Database storage of templates. Nevertheless, it's important to highlight that PCA has its constraints, such as difficulties in precisely estimating covariance matrices and its incapability to capture inherent invariances unless explicitly incorporated in the training data."

"In a study titled "Emotion Recognition via Sequences," written by A. Mohamed, G.E. Dahl, and G. Hinton, they employed a PCA AAM-based approach to analyze image sequences obtained from the FG-NET consortium for the task of emotion recognition. The findings showcase impressive performance, achieving accuracy rates of 100% for recognizing expressions from extracted faces, 88% for expression recognition from individual frames, and 88% for combined recognition." The authors suggested that it is essential to minimize computational time and complexity to enhance efficiency. Furthermore, they proposed extending these efforts to detecting facial expressions from 3D

photographs, although this approach is costlier than 2D projection and can pose challenges in environments with fluorescent lighting. Additionally, the use of holographic projection in design elements is expensive, and the creation of visuals with 3D holograms is a time-consuming process.

HAAR and AdaBoost were employed for the purpose of emotion recognition, as described in the paper titled "Emotion recognition using HAAR and AdaBoost" by G. Sivaram and H. Hermansky. The study focused on utilizing a strategy based on the Gabor SVM approach, specifically HAAR AdaBoost, using data from the Cohn-Kanade database. Notably, they achieved an accuracy of 89.54% for Mouth Action Units (AUs) in the GS category and 62.81% for H A. "Researchers noticed that the Haar AdaBoost technique demonstrated comparable performance to the Gabor+SVM method for AUs associated with the eye and brow areas but showed significant underperformance when it came to AUs linked to the mouth."

An advantage of their work is the intention to develop a comprehensive, publicly accessible AU database featuring AUs occurring once to support future research endeavors. However, it's important to acknowledge that this approach is sensitive to outliers as each classifier is tasked with correcting the errors of its predecessors, making it heavily reliant on outlier correction. Furthermore, scaling up this process is deemed nearly unfeasible, constituting a notable drawback

Description of the Existing System

Using CNN, recognize a face expression from a provided real-time video and extract face expressions such as nose, eyes and mouth from the discovered input face, as well as split the detected facial emotions into multiple categories such as joyful, sorrow, rage, contempt, and fear. One of the special cases of object detection is face detection. To keep the face of the supplied image, it uses lighting compensation techniques and morphological processes.

The Convolutional Neural Network (CNN) has witnessed substantial advancements in recent times, emerging as a prominent gem within the flourishing domain of deep neural networks. Moreover, computer vision technology empowers artificial intelligence with the capability to perceive and comprehend visual information. Thanks to enhancements in computer hardware performance and the development of extensive image annotation datasets, deep learning-driven computer vision algorithms have achieved remarkable triumphs in traditional computer vision tasks like image categorization, object identification, and image segmentation in recent years.

Object detection has found extensive application in real-world scenarios, including video fire detection, autonomous driving, security surveillance, and UAV scene analysis. It has also been a subject of significant academic research. Presently, object detection algorithms can be categorized into two main groups: classical methods rooted in image processing and detecting object in an image technique based on Convolutional Neural Networks (CNNs).

Building upon this foundation, Girshick et al. introduced R-CNN in 2014, marking a pivotal moment. For the first time, Convolutional Neural Networks were employed to identify objects, resulting in a notable 30% improvement in detection accuracy compared to conventional methods. This development garnered considerable attention and acclaim.

In line with current academic research and practical implementations, CNN-based object detection algorithms exhibit higher accuracy and reduced testing times when compared to their traditional counterparts. As a result, they have almost entirely supplanted the traditional methodology in the field.

Convolutional Neural Network

"Figure 1 provides a visual representation of a typical Convolutional Neural Network (CNN). CNNs comprise several essential elements: an input layer, a convolutional layer, a ReLU layer, a pooling layer, and a fully connected layer, which functions much like the

fully connected layer found in conventional neural networks. A complete CNN is formed by arranging and layering these components together. In practical use cases, the convolutional layer and the ReLU layer are frequently combined and collectively called the convolutional layer. In this configuration, the convolutional layer includes an activation function applied after the convolution operation."

"Both the convolutional layer and the fully connected layer apply transformations to the input data using multiple parameters, which include neuron weights (w) and biases (b), alongside an activation function. In contrast, the ReLU layer and the pooling layer perform fixed-function operations. Throughout the training process, the parameters within the convolutional and fully connected layers are modified to minimize the gradient, ensuring that the CNN's classification scores closely match the labels assigned to each image in the training dataset."

Convolutional neural networks are characterized by three important concepts: local receptive fields, sparse weights, and parameter sharing. These principles contribute to the CNN's superior translation and scale invariance compared to other neural networks, rendering them highly suitable for tasks involving image data.

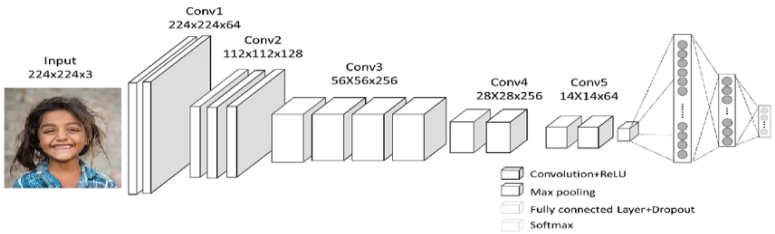

Figure 1: Architecture of CNN Framework

Convolutional Layer

The layer termed as convolution within a convolutional neural network is the most important layer, responsible for the majority of the network's computations. It's essential to clarify that the computation load isn't solely determined by the number of parameters. The convolution technique efficiently lessen the training difficulty of the neural network, along with the network's connections and parameter weights. This makes it easier to train compared to an equivalently scaled fully connected network. There are various common convolution techniques, including regular convolution, transposed convolution, dilated convolution, and depth-wise separable convolution. Regular convolution involves sliding a convolution kernel across the image, processing all pixels in an image with help of a series of matrix computations to analyze their grayscale values. Transposed convolution, conversely, operates from low-dimensional feature mapping to high-dimensional feature mapping, finding applications in areas like semantic analysis and image recognition. Fractionally-Strided Convolution enhances input feature sampling by reducing the transposed convolution's step size while increasing the feature dimension. Dilated convolution, also known as fully convolution , expands the unit's receptive field without increasing parameter count. Depth-wise Separable Convolution is utilized in lightweight network architectures like MobileNets. "It utilizes a single filter for each input channel, followed by pointwise convolution to combine the results obtained from different depth convolution operations." This approach, as opposed to typical convolution techniques, allows for channel and spatial separation, significantly reducing computational load and model size.

Activation Layer

Artificial neural networks leverage activation functions to acquire intricate data patterns. These activation functions, akin to the neuron-based model in the human brain, determine the information to be transmitted to the next neuron. Commonly employed activation functions include, Randomized LeakyReLU (RReLU), Rectified Linear Unit (ReLU),

Exponential Linear Units (ELU), among others. Among these, ReLU stands out as a fundamental unsaturated activation function. Below is its mathematical representation:

$$f(x)=\max(0, x)$$

Pooling Layer

The concept of the pooling layer was initially introduced in the LeNet paper and was referred to as "Subsample" after the release of the AlexNet paper. It has since become a crucial component of modern convolutional neural networks. Placed between consecutive convolutional layers, the pooling layer serves to mitigate overfitting by reducing the volume of data and associated parameters. In the context of processing images, the main role of the pooling layer is to reduce the size of the image. The pooling layer efficiently achieves this compression by conducting collective statistical operations on specific regions within the image, effectively reducing the matrix size. This reduction has the dual advantage of decreasing the parameters required in the final fully connected layer and accelerating computational processing, all while reducing the convolutional layer's excessive sensitivity to the image's precise location. Common operations performed by the pooling layer include maximum pooling, average pooling, spatial pyramid pooling, and more.

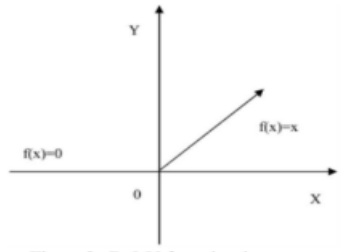

Figure 2: ReLU function image

This technology presents an advanced approach to improve real-time emotion recognition. The fundamental structure of the current system is illustrated in the diagram below, consisting of four main modules. The initial module utilizes a webcam to capture live video feed and employs a local binary patterns (LBP) cascade classifier to detect faces. This detection process involves viewing the image as a collection of small patterns. The subsequent module is focused on image pre-processing, encompassing actions such as cropping, resizing, and intensity normalization. To extract meaningful features from the original image, specific feature extraction algorithms have been developed. Subsequently, the selected features are used for detection. It's worth noting that pre-processing is a critical phase in this system. The following section delves into the pre-processing techniques utilized within the proposed system.

Pre-processing: The webcam image contains non-essential areas for the purpose of facial expression recognition, such as the neck and hair, which do not contribute to the task. Consequently, this irrelevant information has been removed. Otherwise, the detection process would need to contend with additional data, resulting in increased complexity and reduced efficiency. The elimination of this unwanted content from the image is a crucial step in the initial processing of the raw image. Pre-processing steps encompass cropping, resizing, and intensity normalization. Cropping involves trimming the raw image to exclude sections that lack expression-specific details. Emotion detection places particular importance on the facial regions surrounding the mouth and eyes.

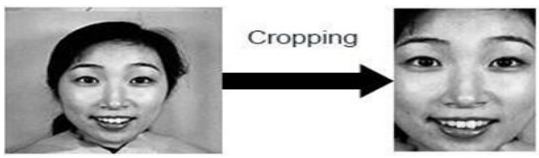

Figure 3: Cropping

Visual brightness and contrast tend to vary based on the illumination and lighting conditions surrounding an object. These fluctuations add intricacy to feature sets and detection techniques. To address these challenges, intensity normalization was implemented. In the proposed method, Min-Max normalization is employed following a linear transformation of the source image. The subsequent section will delve into the feature extraction techniques employed within the system. Data that has been properly pre-processed is fed into the next module, which extracts features.

Feature Extraction

"Feature extraction plays a crucial role in extracting vital information from an image, simplifying the process of emotion detection. In this paper, we employ a hybrid feature extraction method that combines features obtained through Convolutional Neural Networks (CNN) with Histogram of Oriented Gradients (HOG) and facial landmarks. HOG is used to describe the appearance and shape of local objects in an image by analyzing intensity gradients or edge directions. HOG operates on localized cells, making it resilient to geometric variations. These distinctive HOG characteristics vary for each expression, aiding in differentiation. Consequently, HOG was selected as the feature extractor for this system, with HOG features extracted using the skimage module in Python. Following HOG, the subsequent step involves extracting features through facial landmark detection." Facial expression detection is a technique used to identify significant facial features.

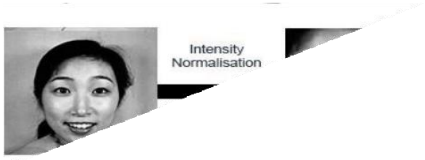

Figure 4: Intensity normalization

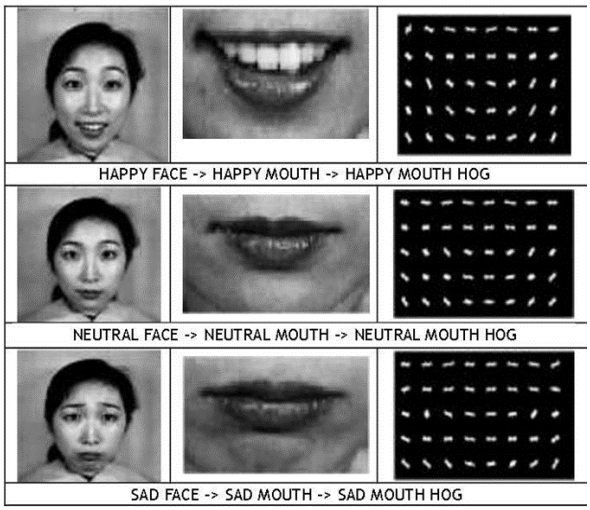

Figure 5: Output after HOG

"The task is accomplished using the Dlib function within OpenCV, which is capable of taking an image region containing an object and generating a set of coordinates that define the object's pose. In this context, it can identify 68 facial landmarks. As mentioned earlier, only the muscles surrounding the lips and eyes are crucial for emotion detection, as these are the muscles that undergo changes during facial expressions. Therefore, only a subset of facial landmarks is taken into account. The results obtained from employing HOG and facial landmark are represented in Figure 4 and 5.

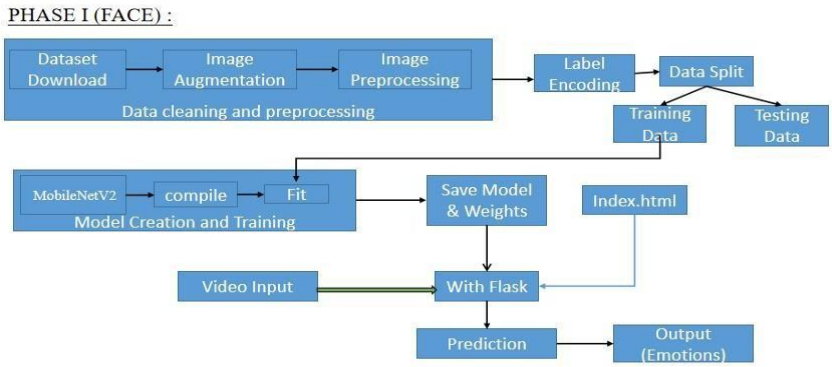

Figure 7: Overall Block Diagram

This system utilizes the features obtained through CNN functionality in combination with the supplementary features extracted in the preceding module to construct a Convolutional Neural Network (CNN) for the purpose of categorizing emotions. The proposed CNN architecture comprises an input layer, four convolutional layers, two pooling layers, and two fully connected classification layers."

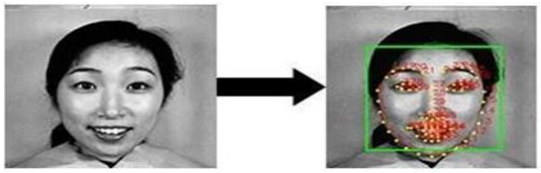

Figure 6: Output after DLib. Facial landmarks are marked in the image

Methodology of the Proposed Work

Since we need to increase the system's recognition rate, extra modifications are made in the third phase utilizing machine learning. Here Static photos and real-time video of a person input can be provided in this way for testing face expressions. This research paper introduces a system designed for facial emotion recognition, aiming to identify human facial expressions encompassing emotions like happiness, sadness, fear, contempt, anger, and surprise. The process begins by capturing a sequence of images from a video containing diverse expressions, saving these images in a database folder.

For each image within this data folder, a mean shape is generated to represent it. The variations in this model, in response to different facial expressions, are measured to calculate the distances or differences between neural and various facial expressions. These computed values are stored in a file directory, and each distinct expression is assigned a unique value for training data purposes. These differences are subsequently employed as input. The system has undergone training to enable it to recognize a range of images and video input sequences portraying various facial emotions.

Block Diagram

"A block diagram is a visual representation that uses labeled blocks to depict individual or multiple items, entities, or concepts, connected by lines to show their relationships. In the context of a system, a block diagram serves as a high-level modularization, breaking down the overall system into distinct, minimally interconnected subsystems. These system block diagrams allow us to view the system as a collection of larger, interlinked components that can be conceptualized and built independently. Block diagrams help in comprehending a system's functions and the connections between its parts. The term 'block diagram' originates from the rectangular blocks employed to represent these components, which are used to describe both hardware and software systems and to illustrate processes."

MobileNet employs depthwise separable convolutions, significantly reducing the number of parameters compared to networks with regular convolutions of the same depth. This

results in the creation of lightweight deep neural networks. Depth wise separable convolutions consist of two operations: depth wise convolution and pointwise convolution. Mobile Net, an open-source CNN class by Google, offers a valuable starting point for training highly efficient and compact classifiers. "MobileNetV2 is composed of two categories of blocks. The initial one is a residual block with a stride of one, while the second is a block with a stride of 2 to reduce dimensions. Both block types comprise three layers. The first layer involves a 1×1 convolution with ReLU6 activation. The second layer is depth wise convolution, and the third layer is a 1×1 convolution without non-linearity. The inclusion of ReLU in this context is crucial for deep networks to preserve the non-zero volume segment of the output domain's strength."

Depth wise Separable Convolution gets its name from the concept of separating the depth and spatial dimensions of a filter. It is defined as a depth wise convolution followed by a pointwise convolution.

The diagram for depth wise separable convolution is given below

Type / Stride	Filter Shape	Input Size
Conv / s2	3 × 3 × 3 × 32	224 × 224 × 3
Conv dw / s1	3 × 3 × 32 dw	112 × 112 × 32
Conv / s1	1 × 1 × 32 × 64	112 × 112 × 32
Conv dw / s2	3 × 3 × 64 dw	112 × 112 × 64
Conv / s1	1 × 1 × 64 × 128	56 × 56 × 64
Conv dw / s1	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 128	56 × 56 × 128
Conv dw / s2	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 256	28 × 28 × 128
Conv dw / s1	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 256	28 × 28 × 256
Conv dw / s2	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 512	14 × 14 × 256
5× Conv dw / s1	3 × 3 × 512 dw	14 × 14 × 512
5× Conv / s1	1 × 1 × 512 × 512	14 × 14 × 512
Conv dw / s2	3 × 3 × 512 dw	14 × 14 × 512
Conv / s1	1 × 1 × 512 × 1024	7 × 7 × 512
Conv dw / s2	3 × 3 × 1024 dw	7 × 7 × 1024
Conv / s1	1 × 1 × 1024 × 1024	7 × 7 × 1024
Avg Pool / s1	Pool 7 × 7	7 × 7 × 1024
FC / s1	1024 × 1000	1 × 1 × 1024
Softmax / s1	Classifier	1 × 1 × 1000

1. Depth wise convolution refers to spatial convolution performed separately for each channel in the input. If we have five channels in the image, we will perform 5 spatial convolutions of size DKxDK.
2. The pointwise convolution with a size of 1x1 is utilized to adjust the dimensionality.
3. Convolution in depth

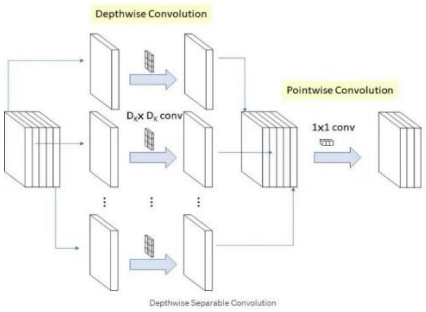

Figure 8: Architecture of Mobilenetv2

This signifies a distinct convolution applied separately to every input channel, leading to an equal number of output channels as there are input channels. This results in a computational expense of.

$$Df^2 * M * Dk^2.$$

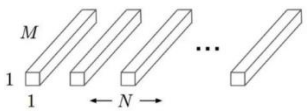

Figure 9: Pointwise convolution

A convolution using a 1x1 kernel that combines the features generated by the depth wise convolution. This incurs a computational cost of...

$$M * N * Df^2.$$

"The differentiation between standard convolution and depth wise separable convolution lies in the design of Mobile Net. Unlike the traditional CNN architecture, Mobile Net splits the convolution process. In conventional CNNs, there's a single 3x3 convolution layer followed by batch normalization and ReLU activation. However, in Mobile Net, this convolution process is separated into two stages: a 3x3 depth-wise convolution and a 1x1 pointwise convolution, as illustrated in the diagram."

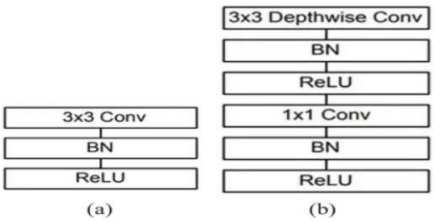

Figure 10: Standard CNN and Mobile net

(a) A conventional convolutional layer with batch normalization and ReLU activation.
(b) Depth-wise separable convolution utilizing depth-wise and point-wise layers, followed by batch normalization and ReLU activation. This device is designed for real-time emotion detection. It aims to reduce computation time. It assists in minimizing multiplication operations during convolution.

Datasets and Data Collection

"The dataset consists of grayscale face images, each sized at 48x48 pixels. These images have been automatically adjusted to have roughly centered faces that occupy a consistent amount of space in each image.

The goal is to categorize each face into one of seven emotion groups based on their facial expressions (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The training dataset comprises 28,709 samples, and the public test dataset includes 3,589 examples.".

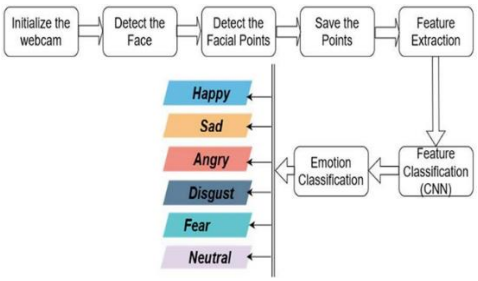

Figure 11: Dynamic FER using CNN Model

Table 1: Image Dataset for 3D-ER CNN Model

Label	No of images
Happy	8989
Sad	6067
Angry	4953
Surprise	4002
Neutral	6198
Disgust	574
Total	30783

The CNN model underwent training using a dataset of 30,395 images that were manually annotated and pre-processed. This dataset was categorized into six classes: neutral/bored, anger, disgust, happiness, sadness, and surprise.

Conclusion

A system capable of reading facial expressions is developed, and a technique in which we can be read in conjunction to develop a more effective method of determining the emotional state of the user and then playing out a music for each state is discovered. This technique can be used in a variety of applications such as interviews, surveys, and so on. This can be improved even further by incorporating more faces to obtain a more precise reading.

References

Jabeen, S.; Mehmood, Z.; Mahmood, T.; Saba, T.; Rehman, A.; Mahmood, M.T. An effective content-based image retrieval technique for image visuals representation based on the bag-of-visual-words model. PLoS ONE 2018, 13, e0194526

Moret-Tatay, C.; Wester, A.G.; Gamermann, D. To Google or not: Differences on how online searches predict names and faces. Mathematics 2020, 8, 1964.

Ubaid, M.T.; Khalil, M.; Khan, M.U.G.; Saba, T.; Rehman, A. Beard and Hair Detection, Segmentation and Changing Color Using Mask R-CNN. In Proceedings of the International Conference on Information Technology and Applications, Dubai, United Arab Emirates, 13–14 November 2021; Springer: Singapore, 2022; pp. 63–73.

Meethongjan, K.; Dzulkifli MRehman, A.; Altameem, A.; Saba, T. An intelligent fused approach for face recognition. J. Intell. Syst. 2013, 22, 197–212. [Google Scholar] [CrossRef]

Elarbi-Boudihir, M.; Rehman, A.; Saba, T. Video motion perception using optimized Gabor filter. Int. J. Phys. Sci. 2011, 6, 2799–2806.

Joudaki, S.; Rehman, A. Dynamic hand gesture recognition of sign language using geometric features learning. Int. J. Comput. Vis. Robot. 2022, 12, 1–16.

Abunadi, I.; Albraikan, A.A.; Alzahrani, J.S.; Eltahir, M.M.; Hilal, A.M.; Eldesouki, M.I.; Motwakel, A.; Yaseen, I. An Automated Glowworm Swarm Optimization with an Inception-Based Deep Convolutional Neural Network for COVID-19 Diagnosis and Classification. Healthcare 2022, 10, 697.

Yar, H.; Hussain, T.; Khan, Z.A.; Koundal, D.; Lee, M.Y.; Baik, S.W. Vision sensor-based real-time fire detection in resource-constrained IoT environments. Comput. Intell. Neurosci. 2021, 2021, 5195508.

Yasin, M.; Cheema, A.R.; Kausar, F. Analysis of Internet Download Manager for collection of digital forensic artefacts. Digit. Investig. 2010, 7, 90–94.

Rehman, A.; Alqahtani, S.; Altameem, A.; Saba, T. Virtual machine security challenges: Case studies. Int. J. Mach. Learn. Cybern. 2014, 5, 729–742.

Afza, F.; Khan, M.A.; Sharif, M.; Kadry, S.; Manogaran, G.; Saba, T.; Ashraf, I.; Damaševičius, R. A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection. Image Vis. Comput. 2021, 106, 104090.

Rehman, A.; Khan, M.A.; Saba, T.; Mehmood, Z.; Tariq, U.; Ayesha, N. Microscopic brain tumor detection and classification using 3D CNN and feature selection architecture. Microsc. Res. Tech. 2021, 84, 133–149.

Haji, M.S.; Alkawaz, M.H.; Rehman, A.; Saba, T. Content-based image retrieval: A deep look at features prospectus. Int. J. Comput. Vis. Robot. 2019, 9, 14–38.

Alkawaz, M.H.; Mohamad, D.; Rehman, A.; Basori, A.H. Facial animations: Future research directions & challenges. 3D Res. 2014, 5, 12.

Saleem, S.; Khan, M.; Ghani, U.; Saba, T.; Abunadi, I.; Rehman, A.; Bahaj, S.A. Efficient facial recognition authentication using edge and density variant sketch generator. CMC- Comput. Mater. Contin. 2022, 70, 505–521.

Rahim, M.S.M.; Rad, A.E.; Rehman, A.; Altameem, A. Extreme facial expressions classification based on reality parameters. 3D Res. 2014, 5, 22.

Rashid, M.; Khan, M.A.; Alhaisoni, M.; Wang, S.H.; Naqvi, S.R.; Rehman, A.; Saba, T. A sustainable deep learning framework for object recognition using multi-layers deep features fusion and selection. Sustainability 2020, 12, 5037.

Lung, J.W.J.; Salam, M.S.H.; Rehman, A.; Rahim, M.S.M.; Saba, T. Fuzzy phoneme classification using multi-speaker vocal tract length normalization. IETE Tech. Rev. 2014, 31, 128–136.

Amin, J.; Sharif, M.; Raza, M.; Saba, T.; Sial, R.; Shad, S.A. Brain tumor detection: A long short-term memory (LSTM)-based learning model. Neural Comput. Appl. 2020, 32, 15965–15973.

Kołakowska, A. A review of emotion recognition methods based on keystroke dynamics and mouse movements. In Proceedings of the 6th IEEE International Conference on Human System Interactions (HSI), Sopot, Poland, 6–8 June 2013; pp. 548–555.